(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number : **0 646 858 A1**

(12) EUROPEAN PATENT APPLICATION

(21) Application number : **94306322.2**

(22) Date of filing : **26.08.94**

(51) Int. Cl.[6] : **G06F 3/06**, G06F 13/40, G06F 11/20

(30) Priority : **07.09.93 US 124653**

(43) Date of publication of application : **05.04.95 Bulletin 95/14**

(84) Designated Contracting States : **DE FR GB**

(71) Applicant : **AT & T GLOBAL INFORMATION SOLUTIONS INTERNATIONAL INC.**
1700 South Patterson Boulevard
Dayton, Ohio 45479 (US)

(72) Inventor : **DuLac, Keith Bernard**
8652 Hila
Derby, KS 67037 (US)

(74) Representative : **Robinson, Robert George**
International Patent Department,
AT&T GIS Limited,
915 High Road,
North Finchley
London N12 8QJ (GB)

(54) Data storage system architecture.

(57) A data storage system comprises a matrix of intelligent storage nodes interconnected to communicate with each other via a network of busses (R_o-R_m,C_o-C_n). The network of busses includes a plurality of first busses (R_o-R_m) for conducting data from and to a corresponding plurality of host system processors (H_o-H_m) and a plurality of second busses (C_o-C_n), each one of the second busses intersecting with each one of the first busses. The nodes are located at each intersection. The storage nodes each include a data storage device (D), such as a magnetic disk drive unit, a processor (P) and buffer memory (B1-B3), whereby the node processor controls the storage and retrieval of data at the node as well as being capable of co-ordinating the storage and retrieval of data at other nodes within the network.

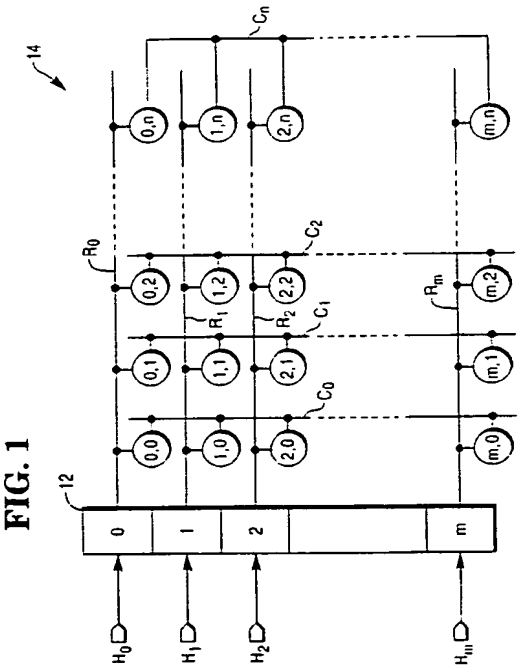


FIG. 1

EP 0 646 858 A1

This invention relates to data storage systems.

Disk array storage systems are known, which include a plurality of disk drives which operate in parallel and appear to the host system as a single large disk drive. Numerous disk array design alternatives are possible, incorporating a few to many disk drives. Several array alternatives, each possessing different attributes, benefits and shortcomings, are presented in an article titled "A Case for Redundant Arrays of Inexpensive Disks (RAID)" by David A. Patterson, Garth Gibson and Randy H. Katz; University of California Report No. UCB/CSD 87/391, December 1987. This article discusses disk arrays and the improvements in performance, reliability, power consumption and scalability that disk arrays provide in comparison to single large magnetic disks. Five different storage configurations are discussed, referred to as RAID levels 1 to 5, respectively.

Complex storage management techniques are required in order to coordinate the operation of the multitude of data storage devices within an array to perform read and write functions, parity generation and checking, data restoration and reconstruction, and other necessary or optional operations. Array operation can be managed by a dedicated hardware controller constructed to control array operations. However, the data storage performance achievable using this technique is limited.

It is an object of the present invention to provide a data storage system which enables a high data storage performance to be achieved.

Therefore, according to one aspect of the present invention, there is provided a data storage system including a network of nodes, interconnected to communicate with each other via a plurality of busses, characterized in that each one of said nodes includes: a data storage device connected to receive data from and provide data to at least one of said plurality of busses; and a node processor connected to said at least one of said busses for controlling the storage and retrieval of data at said data storage device, said node processor being capable of controlling the storage and retrieval of data at data storage devices associated with additional nodes of said plurality of nodes through communications between said node processor and said additional nodes via said plurality of busse.

According to another aspect of the present invention, there is provided a method for transferring data between a host processor and a matrix of data storage nodes, each node including a data storage device and control logic for coordinating data storage operations for a plurality of data storage nodes, characterized by the step of selecting any one of said data storage nodes to control the transfer of data between said host processor and a first subset of said matrix of data storage nodes.

One embodiment of the present invention will now be described by way of example, with reference to the accompanying drawings in which:-

Figure 1 is a diagrammatic illustration of a data storage system including a plurality of disk drives and inexpensive processors located within a matrix network, constructed in accordance with the present invention; and

Figure 2 is a block diagram showing the processor, disk drive, and associated elements located within each node of the matrix network illustrated in Figure 1.

Referring now to Figures 1 and 2, there is seen a data storage system in accordance with the present invention. The architecture shown in Figure 1 includes a host processor connection block 12 providing connection to one or more host system processors, not shown. The host processors are identified by reference numerals H_0, H_1, H_2, ···H_m. Connection block 12 couples host processors H_0, H_1, H_2,···H_m to a network 14 of data storage nodes. Network 14 includes several busses, R_0 through R_m, arranged in rows, each bus connecting one of host processors H_0 through H_m with a group of storage nodes. Network 14 further includes several busses, C_0 through C_n, arranged in columns. A node is formed at every intersection between a row and column bus. The nodes are identified by pairs of coordinates, the first coordinate referring to the number of the row bus to which it connects, and the second coordinate identifying the column bus to which the node connects. The network includes nodes from (0, 0), at the intersection of busses R_0 and C_0, through (m, n), at the intersection of busses R_m and C_n.

H_0 in the configuration shown is connected directly to storage nodes (0, 0) through (0, n) through bus R0. In addition, H_0 is provided access to all the storage nodes on bus C_0, i.e., nodes (1, 0) through (m, 0) by passing through node (0, 0). Nodes (0, 1) through (0, n) similarly provide access for processor H_0 to the nodes on busses C_1 through C_n, respectfully. Each one of host processors H_1 through H_m has direct access to all the storage nodes on busses R_1 through R_m, respectively, and access through interconnecting nodes to all the storage nodes on network 14.

Host processor connection block 12 may include logic for executing group array algorithms, such as the RAID algorithms that are necessary for issuing I/O operations, handling error exception conditions, and performing data reconstruction, when a storage device in network 14 fails. Other functions of the logic included in connection block 12 may include diagnostic and group algorithm initialization executed in response to input provided by a system administration. In a high performance configuration, a host processor connection block

will exist for every row bus (R_0 through R_m) that is shown in node network 14. The high performance configuration allows multiple I/O commands and data to flow over the attached row busses simultaneously. In a lower performance, lower cost configuration command and data flow over one row bus.

Each of storage nodes (0, 0) through (m, n) includes a storage device, node processor, buffers and interface logic as shown in Figure 2. A block diagram showing the processor, disk drive, and associated elements located within node (m, n) of network 14 is shown.

Node (m, n) is seen to include an interface I/F 1 to column bus C_n, a second interface I/F 2 to row bus R_m, an inexpensive processor P, data buffers B1, B2, I and B3, and a storage element D, such as a Head Disk Assembly (HDA) for storing and retrieving data. Node processor P and data buffers B1, B2, I and B3 are connected to interface I/F 1 and thereby to network bus C_n by a node bus identified as BUS 1. A second bus, identified as BUS 2, provides connection between node processor P and data buffers B1, B2, I and B3 and interface I/F 2, which thereby provides connection to network bus R_m. Read/write buffer B3 also provides the node connection to storage element D. Nodes (0, 0) through (m, n) are similarly constructed.

Node processor P, in a conventional sense, controls the network protocols, buffer management, error recovery and storage media control such as head positioning, data encoding/decoding and defect handling. A typical example of the network node could be a Small Computer System Interface (SCSI) disk drive.

In operation, array storage requests are received from one or more host processors and directed to designated nodes within network 14 for execution. An exemplary array operation could be for H_0 to issue a RAID level 5 write operation. It will be appreciated that in a RAID level 5 configuration, data and parity information are distributed over a plurality of disks. The command is formed in a packetized mode for serial connections, or in handshake mode for parallel connections, and issued to appropriate nodes over bus R_0. H_0 could issue a write to any desired node (0,0) to (0,n) residing on bus R_0 The node that receives the command will be referred to in the discussion which follows as the primary node. Remaining network nodes will be referred to as secondary nodes. The command contains information about secondary nodes that will be involved in subsequent read/write operations emanating from the primary node. These operations are necessary to complete the RAID level 5 write command. The primary node upon receiving a command takes responsibility for the operation if no error conditions occur. The primary node will report status conditions to the appropriate host processors for irregular conditions .

The data storage system described above permits the distribution of the compute power necessary to execute the array algorithms and functions to the nodes of a generalized network. The network can consist of intelligent disk drives such that the array algorithms and most common functions are executed at the array nodes.

The host system is relieved of many of the array storage operations. Additionally, several array requests may be executed concurrently, each request being processed by a different primary node. The system thereby realizes increased performance beyond the capabilities of a storage system employing a single hardware controller.

The two main attributes of the described system are:

1. Increase in performance because each node contains sufficient processor power to relieve either the Host processor or the H/W array processor; and
2. Relieves the bandwidth bottleneck of the I/O connection since multiple I/O paths can be used to connect the array nodes.

The invention, therefore is very adaptable to various network architectures and provides improvements in network storage performance. This is due to the compute power which is available independent of host system application load. The invention is also intended to improve the incremental capacity and the reliability of computer networks.

It is important to note that network 14 can be a generalized switching arrangement that would provide a multitude of paths to the individual storage devices coupled to the network.

Listed below are examples to show the execution of the exemplary operations by the storage system according to the present invention.

Operation Number	Host	Primary Node	Secondary Node	Operation
1	H0	(0,1)	(1,1)	Write
2	H1	(1,0)	-	Read
3	H2	(2,2)	(1,2)	Write

Operation 1:

H_0 issues a RAID level 5 write to node (0,1). H_0 passes commands and data to node (0, 1) processor P and buffer B1, respectively, over network bus R_0 and node bus BUS 1. Node (0, 1) processor P decodes the command and determines a read-modify-write cycle is necessary involving secondary node (1,1). Node (0, 1) processor P issues a read command with node (0, 1) identified as the source to node (1,1). The command is communicated to node (1, 1) via bus C_1.

Simultaneously processor P in node (0,1) issues a read to HDA device D in node (0,1) to read old data from HDA device D into Buffer I.

Node processor P in node (1,1) receives the read command via bus C_1, interface block I/F 1, and node bus BUS 1. Node (1,1) processor P decodes the received read command and retrieves read data from HDA device D into buffer I. Node (0,1) and (1,1) complete their respective reads asynchronously. When the reads are complete, node (0,1) contains new data in buffer B1 and old data in buffer I. Node (1,1) contains old parity in its buffer I. Node (1,1) informs node (0,1) that old parity data is in buffer. Node (0,1) reads old parity data over column bus C_1 into node (0,1) buffer B2. Node (0,1) now has new data, old data and old parity in its buffer.

To complete the RAID 5 write operation, node processor (0,1) orders an exclusive-OR of the data stored within buffers B1, B2 and I to generate the new parity data. The new parity is placed in buffer I and readied for transmission to node (1,1) for parity update. Simultaneously, node (0,1) writes the new data from buffer B1 to buffer B3 for writing to storage device D. Node (0,1) issues a normal write command of new parity from buffer I.

Node (1,1) informs node (0,1) that parity write is complete and, in turn, when node (0,1) completes write of new data, informs host processor H_0 that the RAID level 5 write is complete..

Operation 2:

Host processor H_1issues a normal read to node (1,0) over row bus R_1. Upon completion of the read, node (1,0) reports over bus R_1 to processor H_1 that the operation has completed.

Operation 3:

Operation 3 occurs identical to operation 1 except command and data is passed over row bus R_2 and column bus C_2 and report operation complete messages provided to host H_2 over bus R_2.

Operations 1, 2 and 3 may be performed concurrently.

As shown by the examples described above, the architecture enables multiple concurrent operations that distributes the RAID algorithms over the array of nodes. The nodes act as peers and operate in a dynamic client/server mode. This invention facilitates expansion of nodes in both row and column directions. Such expansion permits improvement in performance and capacity without impacting the host processor performance.

The node operation is generalized and could be implemented so that each node can manage as a primary or secondary mode and communicate over a multiplicity of channels.

It can thus be seen that there has been provided by the present invention a data storage system which provides increased performance beyond the capabilities of a host system managed storage system or a storage system employing a single hardware controller. The system described above permits the execution of multiple storage operations concurrently, each operation being coordinated by a different node within the storage network.

This architecture is scalable by design and may be expanded by the addition of nodes in both the row and column direction. In addition, the architecture is not limited to use with magnetic disk drive devices. It can be used to provide RAID technology on sequential access devices (e.g. QIC tapes, DAT tapes, etc.) as well as other direct access devices (e.g., optical disks and media changers) and robotic media changer storage devices. The system can be connected to a single host processor or may be interconnected with several host processors within a multiple processor computer system.

Claims

1. A data storage system including a network of nodes, interconnected to communicate with each other via a plurality of busses (R_o-R_m;C_o-C_n), characterized in that each one of said nodes includes: a data storage device (D) connected to receive data from and provide data to at least one of said plurality of busses (R_o-R_m;C_o-C_n); and a node processor (P) connected to said at least one of said busses (R_o-R_m; C_o-C_n) for con-

trolling the storage and retrieval of data at said data storage device (D), said node processor (P) being capable of controlling the storage and retrieval of data at data storage devices (D) associated with additional nodes of said plurality of nodes through communications between said node processor (P) and said additional nodes via said plurality of busses (R_o-R_m;C_o-C_n).

2. A data storage system according to claim 1, characterized in that said plurality of busses includes a plurality of first busses (R_o-R_m) adapted to transmit data from and to a corresponding plurality of host system processors (H_o-H_m); and a plurality of second busses (C_o-C_n), each one of said plurality of second busses (Co-Cn) intersecting with each one of said plurality of first busses (R_o-R_m); said nodes being located at the intersections of said first busses (R_o-R_m) with said second busses (C_o-C_n).

3. A data storage system according to claim 2 characterized in that within each one of said nodes: said data storage device (D) is connected to receive data from and provide data to the first and the second busses (e.g. R_m, C_n) associated with said one of said nodes; and in that said node processor (P) is connected to said first and second busses (R_m,C_n) associated with said one of said nodes for controlling the storage and retrieval of data at said data storage device (D).

4. A data storage system according to claim 3, characterized in that each one of said nodes further includes a first buffer (B1) connected between said first and second busses (R_m,C_n) associated with said one of said nodes whereby data transmission between said first and second busses (R_m,C_n) associated with said one of said nodes are conducted through said first buffer (B1).

5. A data storage system according to claim 4, characterized in that each one of said nodes further includes: a second buffer (B3) connected between said first and second busses (R_m,C_n) associated with said one of said nodes, second said buffer (B3) being connected to said storage device (D), whereby data transmission between said data storage device (D) and said first and second busses (R_m,C_n) associated with said one of said nodes is effected through said second buffer (B3).

6. A data storage system according to claim 5, characterized in that said data storage device (D) includes a magnetic disk drive.

7. A method for transferring data between a host processor (H_m) and a matrix of data storage nodes, each node including a data storage device (D) and control logic (P) for coordinating data storage operations for a plurality of data storage nodes, characterized by the step of selecting any one of said data storage nodes to control the transfer of data between said host processor (H_o-H_m) and a first subset of said matrix of data storage nodes.

8. A method according to claim 7, characterised by the step of selecting a second one of said data storage nodes to control the transfer of data between said host processor (H_m) and a second subset of said matrix of data storage nodes, whereby the transfer of data between said host processor (H_m) and said first subset, and the transfer between said host processor (H_m) and said second subset, may be performed concurrently.

FIG. 1

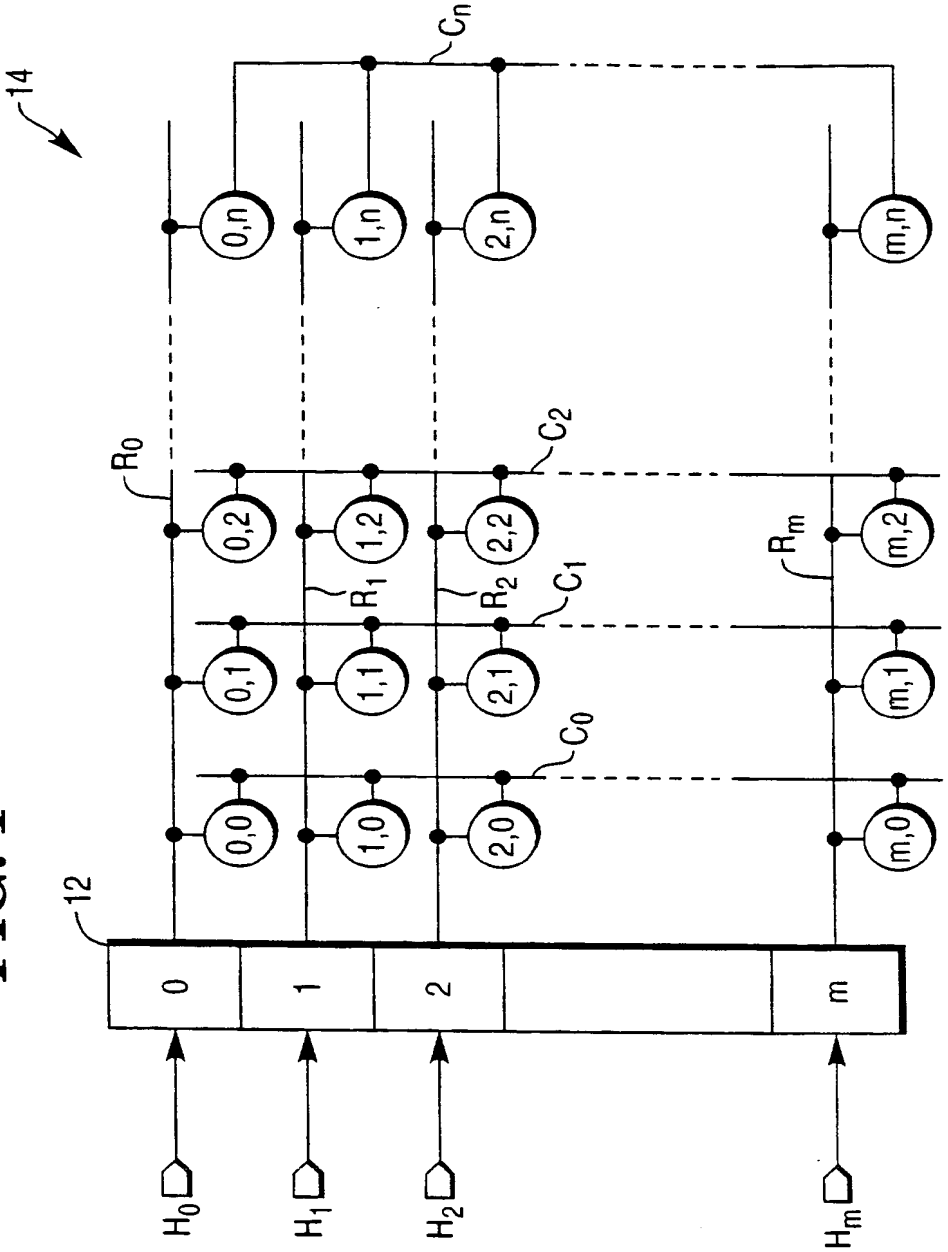
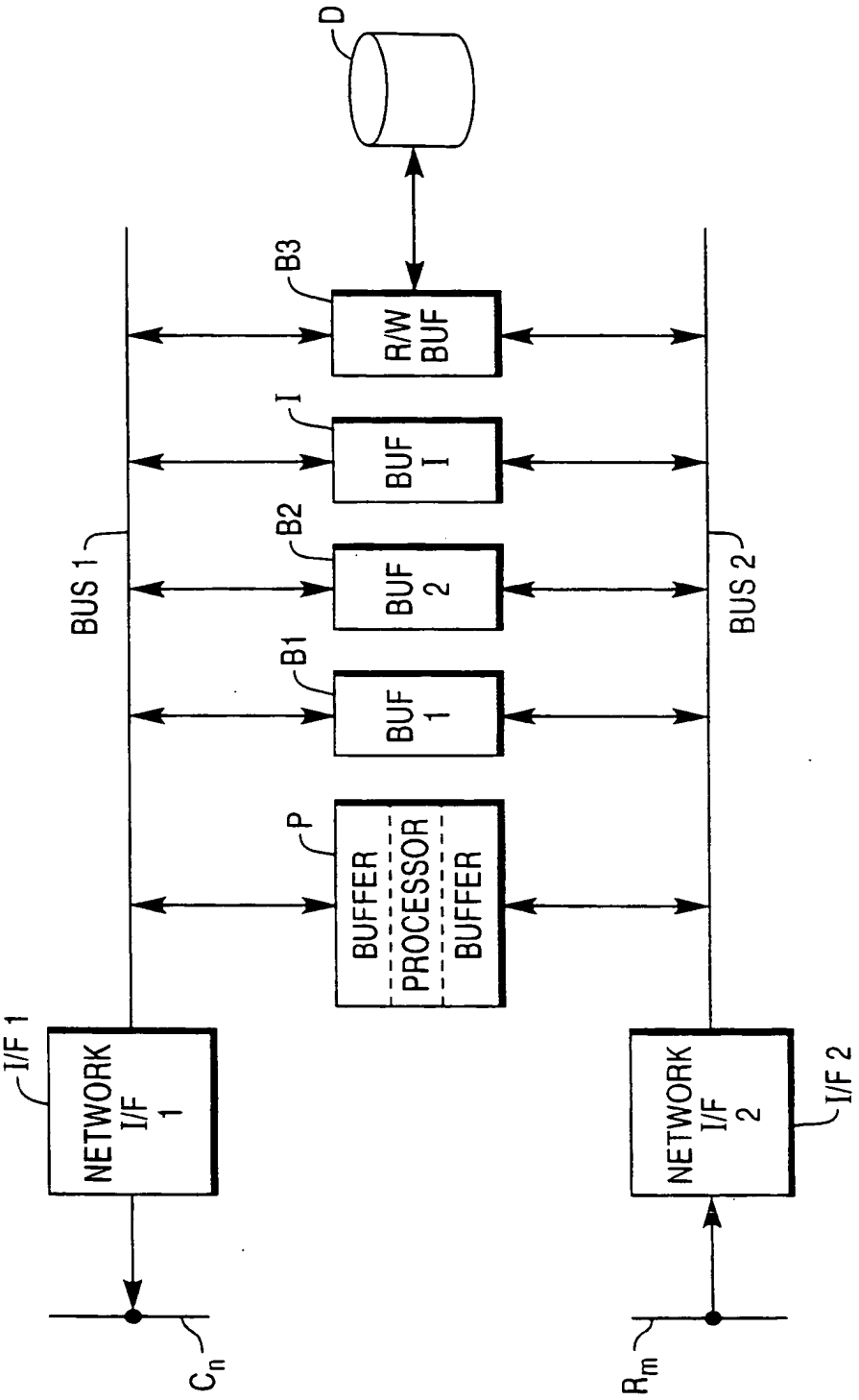


FIG. 2



European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 94 30 6322

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
Y	ELEVENTH IEEE SYMPOSIUM ON MASS STORAGE SYSTEMS - CRISIS IN MASS STORAGE, October 1991, MONTEREY, CA, US pages 131 - 136 J. WILKES 'DataMesh - PARALLEL STORAGE SYSTEMS FOR THE 1990s' * page 131, left column, paragraph 1 - page 134, left column, paragraph 4 *	1-8	G06F3/06 G06F13/40 G06F11/20
Y	US-A-4 807 184 (C. F. SHELOR) * column 2, line 48 - column 3, line 54; figure 1 *	1-8	
A	COMPUTER DESIGN, vol.27, no.20, 1 November 1988, WESTFORD, MA, US pages 23 - 25 D. LIEBERMAN 'Parallel machine extends scalability to encompass I/O processing' * Lower half of page 24, columns 1-3 * * page 25, column 1, paragraph 2 - column 3, paragraph 1; figure *	1-8	TECHNICAL FIELDS SEARCHED (Int.Cl.6) G06F
A	ELECTRONICS, vol.62, no.2, February 1989, HASBROUCK HEIGHTS, NJ, US pages 97 - 100 T. MANUEL 'BREAKING THE DATA-RATE LOGJAM WITH ARRAYS OF SMALL DISK DRIVES' * page 98; figure * * page 100, column 1, paragraph 4 -paragraph 5 *	1-8	
A	EP-A-0 550 853 (MITSUBISHI DENKI KABUSHIKI KAISHA) * the whole document *		

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
BERLIN	20 December 1994	Abram, R

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)